# Mapping Hindi-English Text Re-use Document Pairs

Parth Gupta[1], Khushboo Singhal[2]

[1] Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
http://users.dsic.upv.es/grupos/nle
pgupta@dsic.upv.es
[2] IR-Lab, DA-IICT, India.
http://irlab.daiict.ac.in
khushboo_singhal@daiict.ac.in

**Abstract** In this working note, we present our system developed to find documents which have Cross-Language Text Reuse between Hindi-English language pairs. We try to see the impact of available resources like Bi-lingual Dictionary, WordNet and Transliteration for the specified task. We use Okapi BM25 model to calculate the similarity between document pairs. The best runs stands at $5^{th}$ position in the competition and recall is second highest among all runs.

## 1 Introduction

Here Text Re-use, tries to project the phenomena of 'Plagiarism' where it refers to "the unauthorized use or close imitation of the language and thoughts of another author and the representation of them as one's own original work, as by not crediting the author"[1]. With easy access to the information with prolific World Wide Web makes it essential to check the authenticity of the work in certain situations like research papers, dissertations, student reports and so on. Cross-Language Text Re-use is the special case where the information is taken from the the source in different language.

In last years, text re-use detection has attracted Information Retrieval and Natural Language Processing communities and the state-of-the-art is advanced with evaluation campaigns like PAN[2] at Cross-Language Evaluation Forum (CLEF)[3]. Text Re-use system identifies the re-used text fragments in the given suspicious documents, if any, from the source documents available. The Text Re-use detection systems are broadly comprised of 4 steps: 1) pre-processing, which consists of the normalization of text, language identification and/or translation of documents; 2) selection of candidate documents, i.e., the selection of a small subset of a large source documents collection as potential source of text re-use; 3) detailed analysis, which implies the investigation of suspicious and source documents in detail to identify the re-used text sections; and 4) post-processing, which consists of merging the detected parts of a single re-use case, removing detected cases which are properly cited. [7]

[1] http://dictionary.reference.com/browse/plagiarism
[2] http://pan.webis.de
[3] http://clef-campaign.org/

State-of-the-art for Cross-Language text similarity especially catering text re-use comprises of different models like multilingual thesauri based in [1,10], Comparable corpora based strategy in [6], Machine Translation based statistical similarity in [5] and char n-gram matching in [3]. To the best of our knowledge, there does not exist any study between Hindi-English text re-use identification. Therefore, we intend to test the performance of presently available resources between specified language pair and check their potential to contribute for the given problem.

In the present approach, we transform the Hindi documents in English documents comparable space by the means of available resources like bilingual dictionary, WordNet and Transliteration engine. Thereafter, we calculate the similarity based on the Probabilistic Model Okapi BM25 [9]. From our experiments and analysis in [8,2], we believe that statistical word based similarity has an edge over word n-grams because they leverage to match the text sections in handling obfuscation more easily.

The problem statement in CLiTR track was to identify the most potential source document of the text re-use if any in the given suspicious document. The source documents are in the English while the suspicious documents are in Hindi. The system developed to address the aforementioned problem is described in Section 2. We report the results in Section 3 while in Section 4 we present the analysis of the results. Finally in Section 5 we conclude the work and talk about future activities.

## 2 Approach

It is a two phase process where, Phase-1 stands to transform the documents in language of comparison - English, while in Phase-2 the similarity is calculated to find the source document of text re-use. These phases are described below

### 2.1 Phase-1

In order to compare the Hindi suspicious documents with English source documents, we use different Natural Language resources like bilingual dictionary, wordnet and transliteration system. We have tested three different approaches basically differentiated by resources used.

**Bilingual Dictionary (D):** We substitute each term $t_i$ of suspicious document $q$ by its corresponding English dictionary word. We use The Hindi Universal Word (UW) dictionary[4] freely available for research, which contains total 134968 words. If the term does not have dictionary entry, we ignore it.

**Wordnet + Bilingual Dictionary (W+D):** In this method we look for each term $t_i$ in the Hindi Wordnet[4] and retrieve all its senses as well as synonyms from it so the vocabulary $v$ of the suspicious document $q$ increases to $v'$. Now we substitute each term in $v'$ by its English dictionary word to prepare the new suspicious document $q'$. The terms which do not have dictionary entry are ignored.

[4] http://www.cfilt.iitb.ac.in/hdict/webinterface_user/index.php

**Bilingual Dictionary + Transliteration (D+T):** In this method, we look for each term $t_i$ of the suspicious document $q$ in the bilingual dictionary and replace the term if there exist an English term for $t_i$. If $t_i$ does not have dictionary entry then we transliterate the term $t_i$ using Google Transliterate API[5].

### 2.2 Phase-2

After transforming the Hindi suspicious documents into English using above mentioned ways, we calculate the similarity score between each suspicious document and all the source documents. At the core, the algorithm is to find the closest source documents for each suspicious document with the Vector Space Model. Now for each suspicious document $q$, we find the closest source document $s$ from all the source documents set $S$. The distance between the documents is measured in terms of BM25 score which is defined as below

$$BM25\ Score(q,s) = \sum_{i=1}^{n} idf(q_i) * \frac{f(q_i,s) * (k_1+1)}{f(q_i,s) + k_1 * (1 - b + b * \frac{|s|}{avgdl})} \quad (1)$$

where, $q_i$ is $i^{th}$ term in document $q$, $f(q_i, s)$ specifies the frequency of term $q_i$ in document $s$. $|s|$ signifies the length of $s$, while $avgdl$ refers to average length of the documents in the source corpus, $k_1$ and $b$ are constants with value 1.2 and 0.75 respectively. While $idf(q_i)$ represents the Inverse Document Frequency of term $q_i$ which is calculated as below

$$idf(q_i) = \log(\frac{N}{df_{q_i}}) \quad (2)$$

where, N is the total number of documents in the corpus while $df_{q_i}$ is signifies number of documents in which term $q_i$ appears.

In the Phase-2 we introduce a similarity threshold $\theta$. If a suspicious document does not have any source document with similarity score above the threshold, we consider it free from text re-use.

## 3 Results

We tested the above mentioned strategies on the training & test data. Table 1 contains the results on training data.

[5] www.google.com/transliterate

| Method | Precision | Recall | F-Measure |
|---|---|---|---|
| D | 0.4545 | 0.6923 | 0.5488 |
| W+D | 0.1717 | 0.2615 | 0.2073 |
| D+T | 0.5051 | 0.7692 | 0.6097 |

**Table 1.** Results on training data

After looking at the high value of recall we worked on improving the precision. In order to reduce the false positives, we introduced a similarity threshold in Phase-2. Table 2 describes the evaluation performance with different threshold values on training data.

| $\theta$ Value | Precision | Recall | F-Measure |
|---|---|---|---|
| 0.0 | 0.5051 | 0.7692 | 0.6097 |
| 9.0 | 0.5376 | 0.7692 | 0.6329 |
| 10.0 | 0.5371 | 0.7231 | 0.6164 |
| 15.0 | 0.6069 | 0.6769 | 0.6400 |
| 20.0 | 0.6635 | 0.5461 | 0.5991 |

**Table 2.** Effect of similarity threshold on the performance evaluation.

It can be seen that setting the $\theta$ below 9.0 will hurt the precision without gaining in terms of recall and similarly setting it above 20.0 will hurt recall greatly. So between 9.0 and 20.0 based on the empirical tuning we set the $\theta = 15.0$ which achieves the maximum F-Measure on training data.

Table 3 show the results achieved on the test data.

| Method | Precision | Recall | F-Measure |
|---|---|---|---|
| D (Run-1) | 0.342 | 0.580 | 0.430 |
| W+D | NA | NA | NA |
| D+T (Run-2) | 0.474 | 0.804 | 0.596 |
| D+T+$\theta$ (Run-3) | 0.439 | 0.607 | 0.509 |

**Table 3.** Results on test data

## 4 Analysis

From the Table 1 and 3, it is clearly visible that the introduction of transliteration helps. Most of the documents in the corpus are from either Tourism or Computer Science

domains and hence contain a lot of Named Entities. Transliteration helped in identifying such Named Entities without let them be Out Of Vocabulary (OOV) words.

We are surprised to see the performance evaluation with the Bilingual Dictionary only, which itself could fetch the recall till 0.580. Some of the Hindi words were in their morphological forms of root dictionary words and hence could not find English synonyms. If taken good care, results can further be improved. Morphology analyzer can be employed to take care of it.

Moreover, results suggest that the present way of wordnet incorporation is not a good strategy because it incurs topic drift and the performance is drastically affected as seen in the Tables 1 and 3.

Google Transliterate API acts in the strange way sometimes and has imposed some limits on the number of queries per unit time or destination. So that some of the words were not transliterated properly but essentially transliteration helped a lot in identifying the correct document pairs and responsible for the as high recall as 0.804.

Introduction of threshold $\theta$ helped in training data but could not promise the same results on test data. The threshold is not normalized and hence is not robust across different corpora. We wish to investigate it in future because it is a main element, we believe, which acts behind high precision.

## 5 Conclusion and Future Work

The obtained results suggest that available resources are capable enough in finding the text re-use document pairs for Hindi-English. Transliteration is helps in identifying the Named Entities and contribute to obtain higher recall. If morphology analyzer is incorporated to use the bilingual dictionary, results may further increase.

In future, we wish to work on precision of the system. We also wants to see, how the system performs for different amount and nature of text re-use.

## 6 Acknowledgment

The work of the first author has been partially funded by the European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework.

## References

1. Ceska, Z., Toman, M., Jezek, K.: Multilingual plagiarism detection. In: Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications. pp. 83–92. AIMSA '08, Springer-Verlag, Berlin, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-85776-1_8
2. Gupta, P., Singhal, K., Majumder, P., Rosso, P.: Detection of Paraphrastic Cases of Monolingual and Cross-lingual Plagiarism. In: ICON 2011. Macmillan Publishers, Chennai, India (2011)
3. McNamee, P., Mayfield, J.: Character n-gram tokenization for european language text retrieval. Inf. Retr. 7(1-2), 73–97 (2004)

4. Narayan, D., Chakrabarti, D., Pande, P., Bhattacharyya, P.: An experience in building the indo wordnet - a wordnet for hindi. In: First International Conference on Global WordNet, Mysore, India (2002)
5. Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., Rosso, P.: A statistical approach to crosslingual natural language tasks. J. Algorithms 64(1), 51–60 (2009)
6. Potthast, M., Stein, B., Anderka, M.: A wikipedia-based multilingual retrieval model. In: ECIR. pp. 522–530 (2008)
7. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 1–9. CEUR-WS.org (sep 2009), {http://ceur-ws.org/Vol-502}
8. Rao, S., Gupta, P., Singhal, K., Majumder, P.: External & intrinsic plagiarism detection: Vsm & discourse markers based approach - notebook for pan at clef 2011. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
9. Robertson, S., Spärck Jones, K.: Simple, proven approaches to text retrieval. Tech. Rep. UCAM-CL-TR-356, University of Cambridge, Computer Laboratory (Dec 1994), http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-356.pdf
10. Steinberger, R., Pouliquen, B., Hagman, J.: Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In: CICLing. pp. 415–424 (2002)